PROGRAMA RISE

INFORME DE PROGRESO DE LA INVESTIGACIÓN

Estudiantes Subgraduados

Periodo de enero a mayo 2016

INFORMACIÓN DEL ESTUDIANTE

Nombre del estudiante:	Iván L. Jiménez Ruiz
Número de estudiante:	801-11-3205
Concentración:	Ciencias de Cómputos
Periodo estimado de graduación:	Mayo 2017

INFORMACIÓN DEL MENTOR

Nombre del mentor:	Dr. Humberto Ortiz-Zuazaga
Departamento:	Ciencia de Cómputos
Rango mentor:	Full Professor (PhD)
Correo electrónico mentor:	humberto.ortiz@upr.edu

DIVULGACIÓN DE RESULTADOS

Presentaciones de los resultados del proyecto. (**Favor incluir copia del afiche**):

Modalidad de la presentación: Oral ___ Afiche X

Título: Evaluating Quantification and Expression methods with *Drosophila melanogaster* data

Nombre y lugar de la actividad: 11th Annual Research Initiative for Scientific Enhancement (RISE) Area Conference, March 2016. San Juan, PR. [Poster #11]

Fecha: Marzo 2016

Autores: Jimenez-Ruiz, Ivan[1] ; Gonzalez Mendez, Ricardo[2] ; Ropelewski, Alexander[3] ; Agosto Rivera, José[1] ; Ortiz-Zuazaga, Humberto[1]

[1.]University of Puerto Rico, Rio Piedras Campus, San Juan, Puerto Rico

[2.]University of Puerto Rico, Medical Sciences Campus, San Juan, Puerto Rico

Taller de ética y OPASO (BIOL4995 – Investigación Subgraduada) ; Verano 2014

Título del proyecto (inglés):

Evaluating Quantification and Expression methods with
Drosophila Melanogaster data

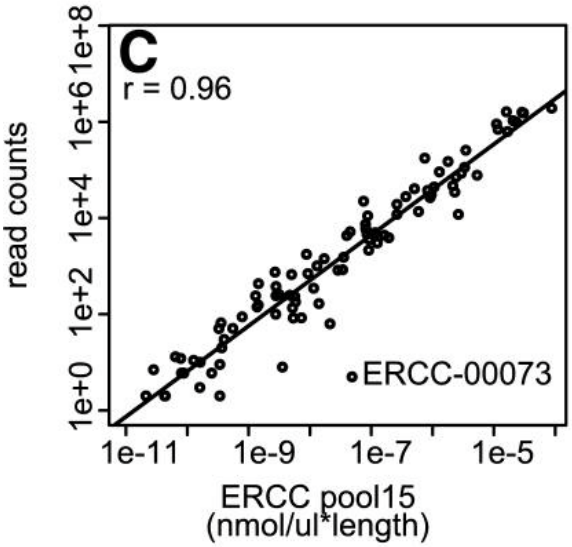
En el siguiente apartado redacte un informe detallado, en inglés, de lo que usted trabajó en su proyecto de investigación durante este semestre. Recuerde incluir aspectos tales como:

- Estado de situación del proyecto
- Resultados obtenidos en el período informado, incluyendo tablas y figuras con sus respectivas leyendas.
- Etapas futuras a desarrollarse
- Dificultades encontradas

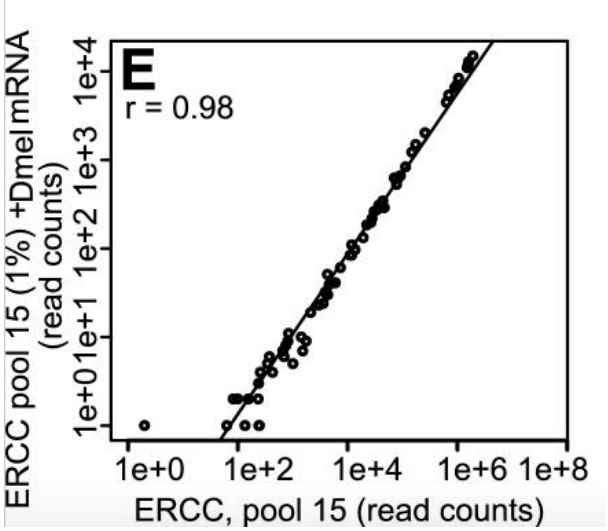
Project Status

The purpose of this project was to compare and contrast de novo and reference-based assembly of RNA sequenced data. To accomplish this, we analyze the data using diverse tools, all of which are presented in the workflow section below (page 12). At the end of the 2015 internship at the Pittsburgh Supercomputing Center, we discovered that there were unidentified viral sequences that annotated to Drosophila Birnavirus and X virus in the original data files. To prove this finding, we have attempted to reproduce the scatter plots C and E (Jiang et al., 2011) found below. These plots refer to ERCC (External Read Control Consortium) genes (see link on page 8) and provide a standard goal we want to achieve with our results from both assemblies. In both cases, we created an ERCC reference using all of the ERCC genes referenced in the paper. We hope the reproduction of these reference-based assembly results will serve as our control to compare against plots generated with a de novo approach. Our hypothesis is that running a de novo based assembly would be just as a powerful as a reference-based assembly at detecting genes that are differentially expressed in an RNA-seq experiment.

Plot 1



Plot 2



Plots 1 and 2. Gene concentrations plotted against read counts for each ERCC (control gene) found after a reference-based assembly of data file SRR039936. On plot 2, the same data file's read counts were plotted against the counts found for data file SRR039935. Links to where one can obtain these data files is provided in the project page. These and other plots from Figure 1 of the paper can be found in at their webpage: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3166838/figure/F1/.

It is important to note that the benefits of validating de novo based assembly procedures include the production of heatmaps (see page 12) and Basic Local Alignment Search Tool (BLAST) verifications. An example BLAST result table is included on page 13.

- Heatmaps:
  - Creates a list of genes that are differentially expressed
  - Provides color-based visualization of gene expression per data file
  - Small phylogenetic tree of listed genes

- BLAST searching:
  - Attempts to predict the origin of RNA sequences (across different species)

Results

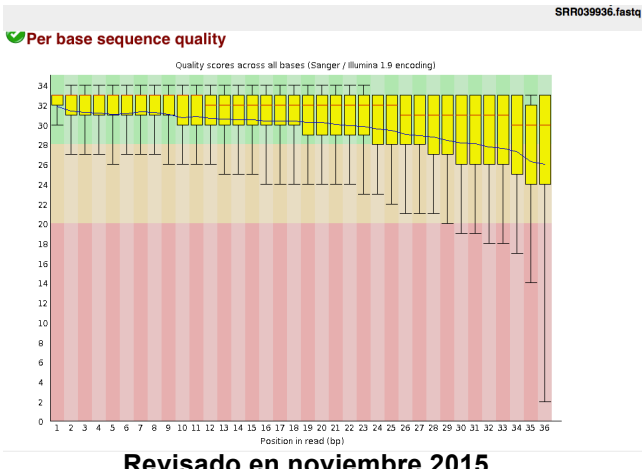
We have developed a simple workflow that encompasses our methodology (see page 11). In summary, we followed these steps:

- Acquire the data files (including D. melanogaster genome and ERCC genome)
- Improve the quality of the data
- Assemble the data (the following steps were done for both reference-based assembly and de novo assembly)
- Quantify: produce counts as RPKMs (Reads per Kilobase of Million mapped reads)
- List differentially expressed genes and represent results via scatter plots

Improved the data before re-running the assemblies.

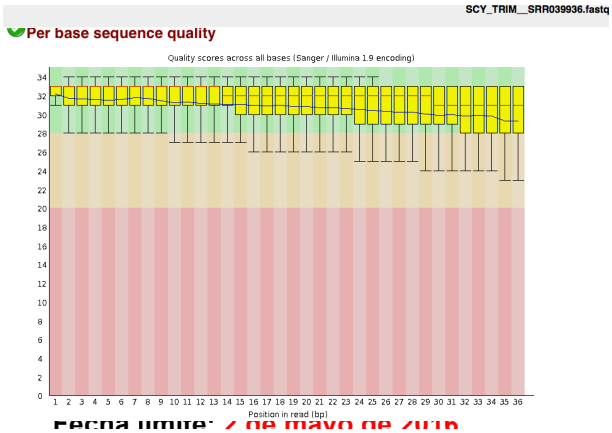
To avoid reevaluate our results from the Summer 2015 internship and move forward, we reacquired our data files from the NCBI database and re-checked the quality of the reads for all files downloaded. Below are two tables that serve as an example of quality improvement using the Scythe, Sickle and FASTQC programs:

Table 1



Revisado en noviembre 2015

Table 2



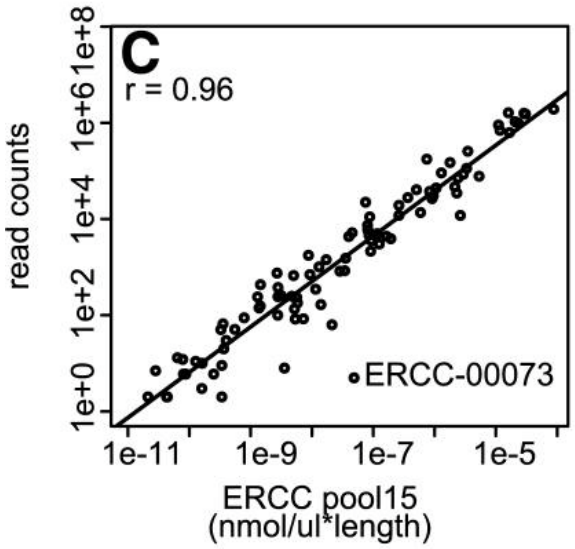
Fecha límite: 4 de mayo de 2016

Tables 1 and 2. Before and after representations of quality scores by base pairs of a single data file (SRR039936.fastq). The quality of the reads in this data file was improved using the Scythe and Sickle programs. Scythe was used to remove adapters that are known to cause problems in future steps of an assembly. Sickle was used to trim the edges off of the millions of reads that make up the data. In Table 1, the quality score (format: phred) per base pair appears to have an exponential decrease near the end edge of the reads. This problem appears to have been corrected in Table 2, grouping at around a quality score of 28+.

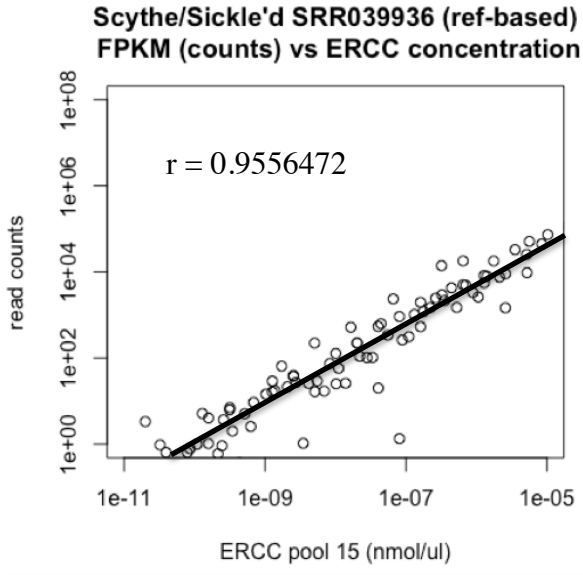
Production of plots for reference-based assembly.

Our goal was to create a script that would output the correct counts of each ERCC gene detected and represented in Plot 1. To ease comparison, Plot 1 has been copied below without its description (see page 4 for description). The scatter plot (Plot 3) can be found below:

Plot 1



Plot 3



Plot 3. Gene concentrations plotted against read counts for each ERCC (control gene) found after a reference-based assembly of data file SRR039936. This data file was sequenced and tagged as consisting of %100 ERCC. The program versions used to produce these counts include TopHat v1.0.13 with Bowtie v0.11.3 and Cufflinks v2.2.1. Each dot represents a gene that aligned and was differentially expressed, with its the corresponding concentration marking the x-axis. Finally, the correlation coefficient of both Plot 1 and 3 appears to be almost identical.

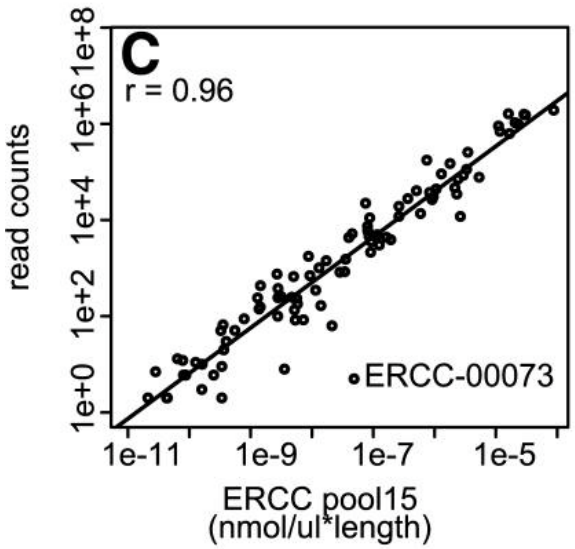
This reproduction is the most important aspect of establishing our control. In Plot 1, the x-axis consists only of concentration: nmol/ul. To explain this, we must take a closer look at Plot 1. If one were to multiply by the length of the sequence that aligned to the ERCC genome: taking ERCC-00073 in Plot 1 as an example, its concentration equals 8.09E-8 and it's length = 603 (http://genome.cshlp.org/content/suppl/2011/07/18/gr.121095.111.DC1/Jiang_TableS1.xls). Concentration times length for this gene equals 0.0000487827 = 4.9E-5, which is not the same as what they shown in the x-axis of Plot 1: ~1E-7. This inconsistency does not take away from the fact that the apparent clustering of dots in Plot 3 begins at about 1E-10 instead of 1E-11 as in Plot 1. Also, the line that is formed by the clustered points in our reproduction appears to skew towards the x-axis. However, Plot 3 represents results that support a correct reference based assembly run.

Fine-tuning of de novo scripts to produce plots that coincide with previous results.

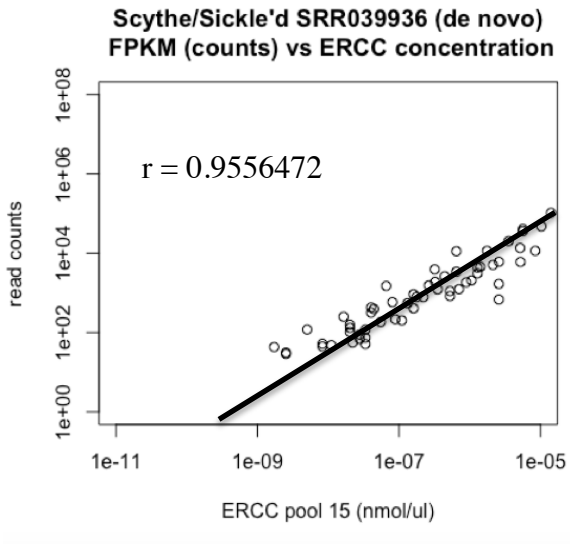
The de novo script that would output the correct counts of each ERCC gene detected and represented in Plot 1 is prone to irregularities by nature. As changing a single parameter for a de novo run can have drastic changes on the output of the programs, further analysis of these approaches is required before publication.

To ease comparison, Plot 1 has been copied below without its description (see page 3 for description). The scatter plot (Plot 4) can be found below:

Plot 1



Plot 4



Plot 4. Gene concentrations plotted against read counts for each ERCC (control gene) found after a de novo assembly of data file SRR039936. This data file was sequenced and tagged as consisting of %100 ERCC. The program versions used to produce these counts include Trinity v2.2.0 RSEM v1.2.30 with Bowtie v0.11.3. Each dot represents a gene that aligned and was differentially expressed, with its the corresponding concentration marking the x-axis. Finally, the correlation coefficient of both Plot 1 and Plot 4 appears to be almost identical.

This reproduction of Plot 1 using a de novo approach follows the same details as in Plot 3, with the x-axis consisting solely of the concentrations of the differentially expressed ERCC genes. As previously mentioned, de novo assemblies are somewhat harder to execute due to the variability in input parameters and their effects on the corresponding output counts. To produce Plot 4, we ran the Trinity program to produce a “make-shift” ERCC genome from our own ERCC reads. Changing a single parameter (i.e. kmer_size = 20) to another value modifies the internal k-mer counting protocol used to produce this experimental genome. This fact implies that there is a greater chance that the counts generated by the alignments of the data reads to the corresponding

"make-shift" genome will vary significantly with each assembly run. As such, it is important to note there is a possibility that the alignments portrayed as counts in Plot 4 are not the best. As seen in the Plot (4), there is a significant amount of dots missing when compared to the results illustrated in Plot 3. This loss of counts at RPKMs below a certain threshold (~50 counts) could have multiple explanations: running an incorrect de novo assembly (either by not setting a parameter correctly or having a default parameter alter the results) or simply that a low-abundance gene could not be reconstructed through this assembly.

Future work

After depicting the benefits of a de novo assembly, the next steps in proving our hypothesis include:

- Modifying the scripts used to produce plots 3 and 4 (in order to analyze additional data files).
- Reproducing Plot 2 for both de novo as well as reference based assemblies.
- Possibly including additional quantification programs for de novo assembly (for example: Oasis, Sailfish)
- Producing heatmaps

The cause for which a cluster of resulting data was lost in Plot 4 needs to be investigated. This includes an in-depth analysis of all parameters (emphasizing default parameters) used by the Trinity program that could influence both the production of the artificial genome as well as the amount of genes that aligned to said genome. In addition, the reproduction of plot 2 requires counts produced by running a reference-based assembly of the data file SRR0399035.fastq against the ERCC genome. Furthermore, the resulting plot could then be compared to the plot made by running an improved de novo assembly on the same data file. re-running de novo based assembly

This project is currently being developed with the goal of publishing all results in a PLOS One paper. Additional steps might include reproducing new heatmaps and performing a BLAST searches as supplemental information. We hope the culmination fo this project will allow us to correctly determine the identity of non-viral, differentially expressed genes in Drosophila melanogaster spike-in data.

REFERENCES

[1] Jiang, Lichun et al. “Synthetic Spike-in Standards for RNA-Seq Experiments.” Genome Research 21.9 (2011): 1543–1551. PMC. Web. First accessed on July 29, 2015.

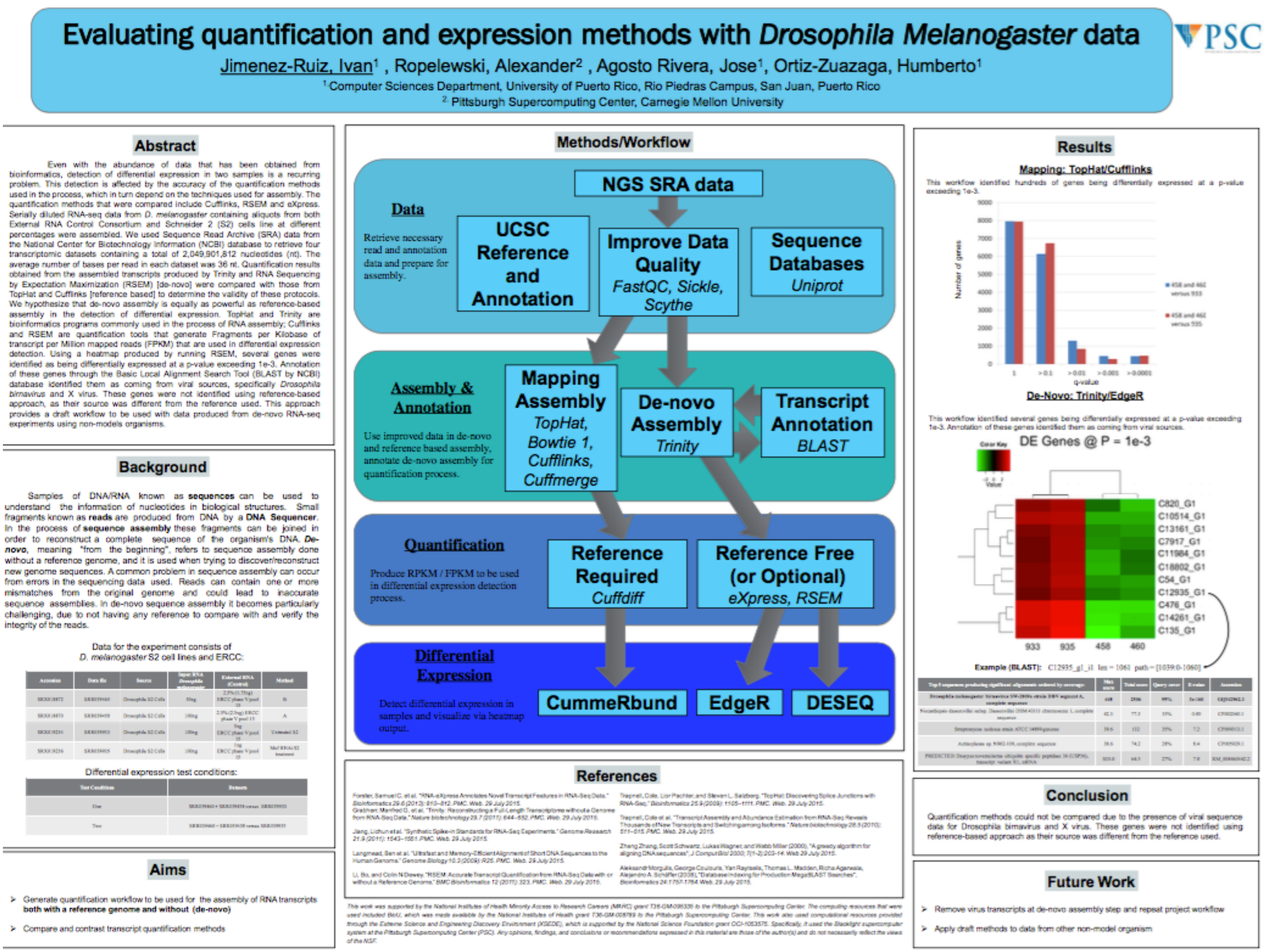
[2] Grabherr, Manfred G. et al. “Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data.” Nature biotechnology 29.7 (2011): 644–652. PMC. Web. 29 July 2015.

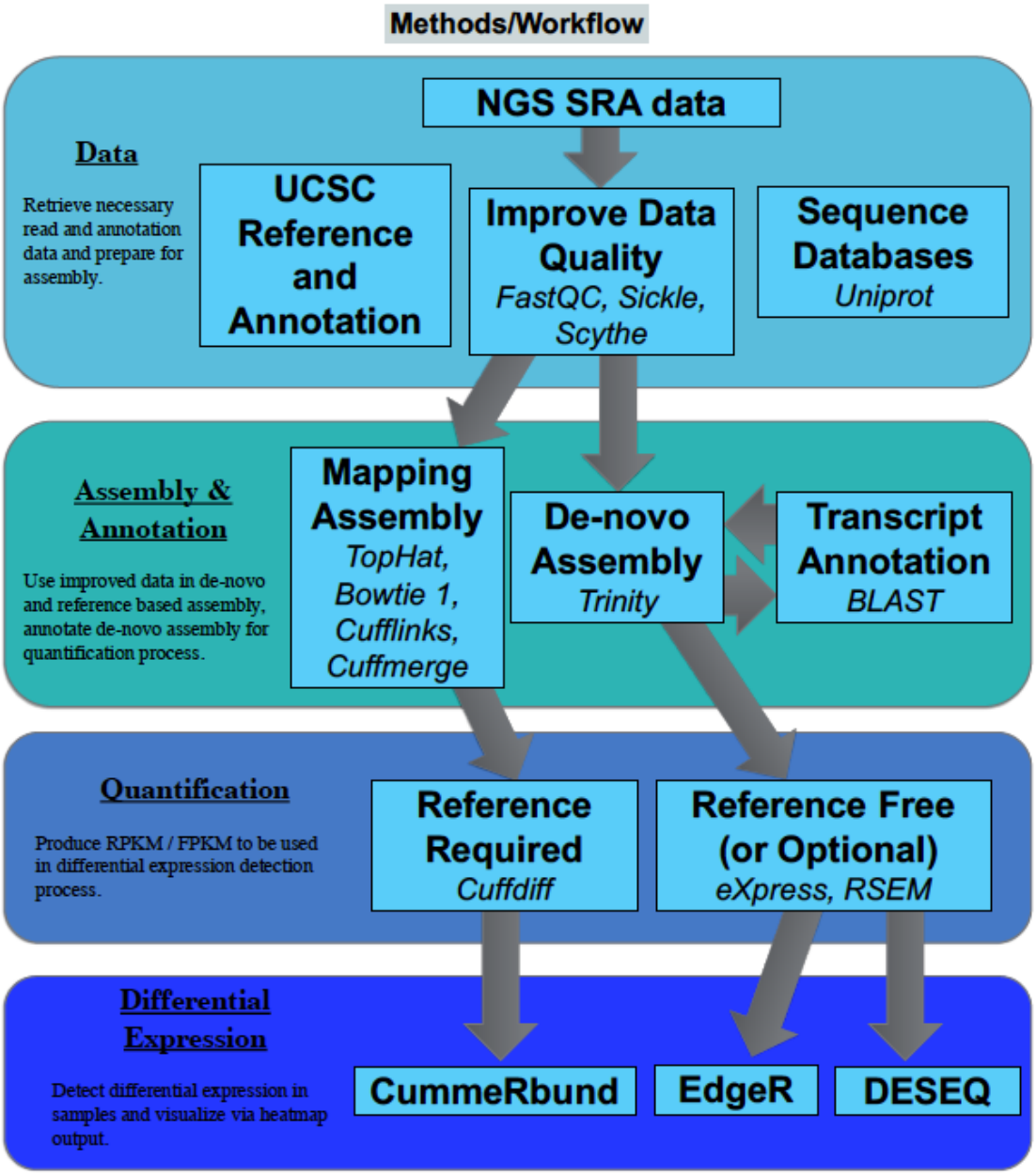
[3] Langmead, Ben et al. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” Genome Biology 10.3 (2009): R25. PMC. Web. 29 July 2015.

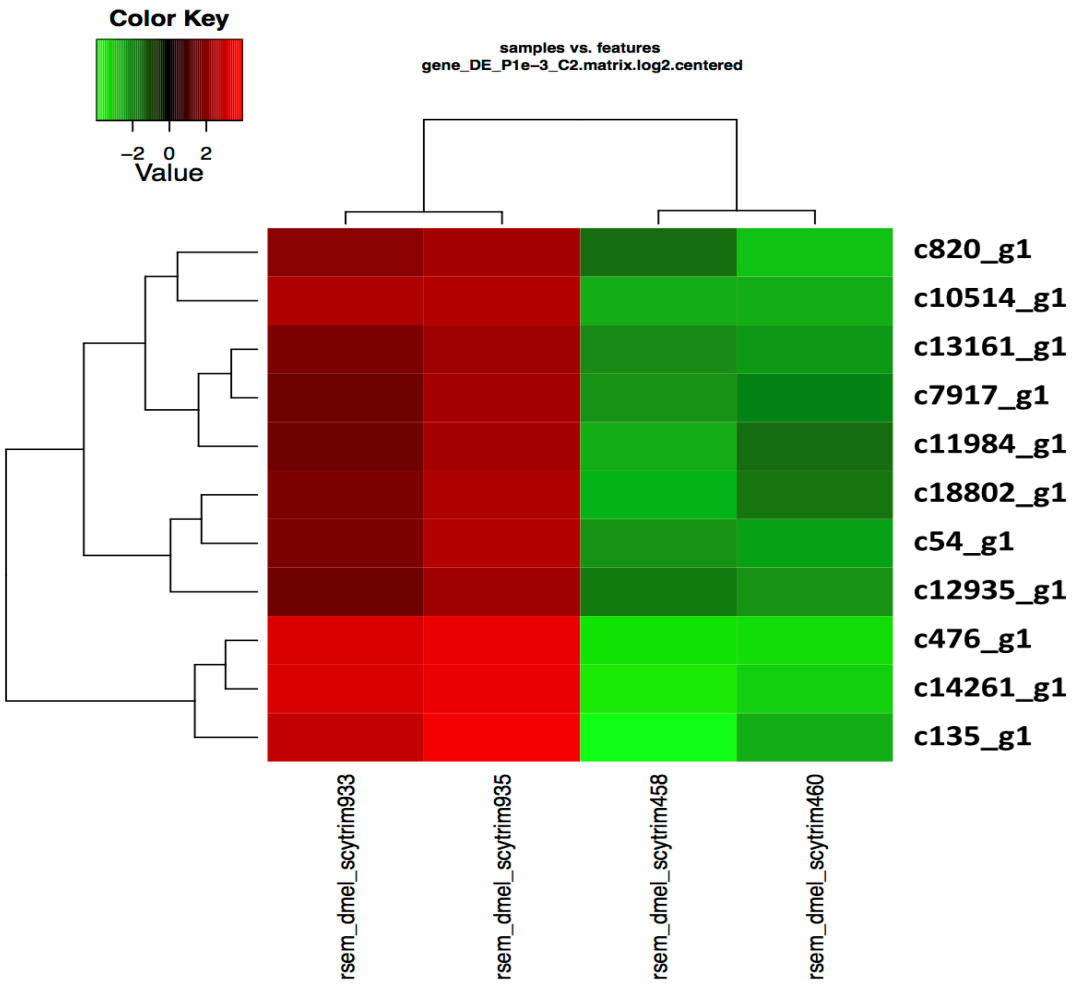
[4] Trapnell, Cole et al. “Transcript Assembly and Abundance Estimation from RNA-Seq Reveals Thousands of New Transcripts and Switching among Isoforms.” Nature biotechnology 28.5 (2010): 511–515. PMC. Web. 29 July 2015.

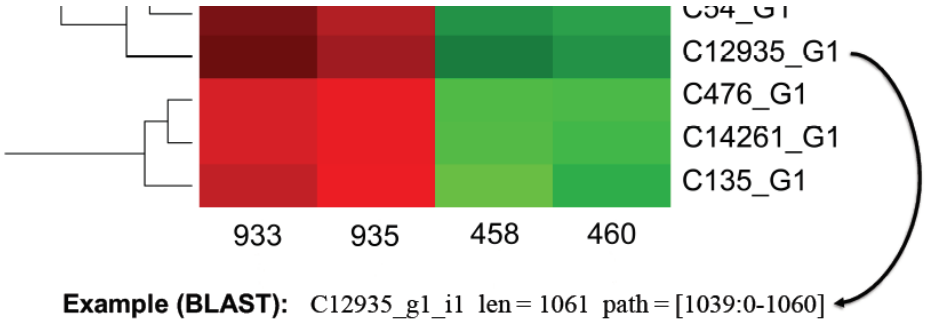
[5] Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. “TopHat: Discovering Splice Junctions with RNA-Seq.” Bioinformatics 25.9 (2009): 1105–1111. PMC. Web. 29 July 2015.

ERCC: http://genome.cshlp.org/content/suppl/2011/07/18/gr.121095.111.DC1/ERCC_genbank_alignments.txt


Evaluating quantification and expression methods with *Drosophila Melanogaster* data
PSC
Jimenez-Ruiz, Ivan[1] , Ropelewski, Alexander[2] , Agosto Rivera, Jose[1], Ortiz-Zuazaga, Humberto[1]
[1] Computer Sciences Department, University of Puerto Rico, Rio Piedras Campus, San Juan, Puerto Rico
[2] Pittsburgh Supercomputing Center, Carnegie Mellon University
Abstract
Even with the abundance of data that has been obtained from bioinformatics, detection of differential expression in two samples is a recurring problem. This detection is affected by the accuracy of the quantification methods used in the process, which in turn depend on the techniques used for assembly. The quantification methods that were compared include Cufflinks, RSEM and eXpress. Serially diluted RNA-seq data from *D. melanogaster* containing aliquots from both External RNA Control Consortium and Schneider 2 (S2) cells line at different percentages were assembled. We used Sequence Read Archive (SRA) data from the National Center for Biotechnology Information (NCBI) database to retrieve four transcriptomic datasets containing a total of 2,049,901,812 nucleotides (nt). The average number of bases per read in each dataset was 36 nt. Quantification results obtained from the assembled transcripts produced by Trinity and RNA Sequencing by Expectation Maximization (RSEM) [de-novo] were compared with those from TopHat and Cufflinks [reference based] to determine the validity of these protocols. We hypothesize that de-novo assembly is equally as powerful as reference-based assembly in the detection of differential expression. TopHat and Trinity are bioinformatics programs commonly used in the process of RNA assembly; Cufflinks and RSEM are quantification tools that generate Fragments per Kilobase of transcript per Million mapped reads (FPKM) that are used in differential expression detection. Using a heatmap produced by running RSEM, several genes were identified as being differentially expressed at a p-value exceeding 1e-3. Annotation of these genes through the Basic Local Alignment Search Tool (BLAST by NCBI) database identified them as coming from viral sources, specifically *Drosophila birnavirus* and X virus. These genes were not identified using reference-based approach, as their source was different from the reference used. This approach provides a draft workflow to be used with data produced from de-novo RNA-seq experiments using non-models organisms.
Background
Samples of DNA/RNA known as **sequences** can be used to understand the information of nucleotides in biological structures. Small fragments known as **reads** are produced from DNA by a **DNA Sequencer**. In the process of **sequence assembly** these fragments can be joined in order to reconstruct a complete sequence of the organism's DNA. ***De-novo***, meaning "from the beginning", refers to sequence assembly done without a reference genome, and it is used when trying to discover/reconstruct new genome sequences. A common problem in sequence assembly can occur from errors in the sequencing data used. Reads can contain one or more mismatches from the original genome and could lead to inaccurate sequence assemblies. In de-novo sequence assembly it becomes particularly challenging, due to not having any reference to compare with and verify the integrity of the reads.
Data for the experiment consists of *D. melanogaster* S2 cell lines and ERCC:
Differential expression test conditions:
Aims
➢ Generate quantification workflow to be used for the assembly of RNA transcripts **both with a reference genome and without (de-novo)**
➢ Compare and contrast transcript quantification methods
Methods/Workflow
NGS SRA data
Data
Retrieve necessary read and annotation data and prepare for assembly.
UCSC Reference and Annotation
Improve Data Quality
FastQC, Sickle, Scythe
Sequence Databases
Uniprot
Assembly & Annotation
Use improved data in de-novo and reference based assembly, annotate de-novo assembly for quantification process.
Mapping Assembly
TopHat, Bowtie 1, Cufflinks, Cuffmerge
De-novo Assembly
Trinity
Transcript Annotation
BLAST
Quantification
Produce RPKM / FPKM to be used in differential expression detection process.
Reference Required
Cuffdiff
Reference Free (or Optional)
eXpress, RSEM
Differential Expression
Detect differential expression in samples and visualize via heatmap output.
CummeRbund
EdgeR
DESEQ
References
Results
Mapping: TopHat/Cufflinks
This workflow identified hundreds of genes being differentially expressed at a p-value exceeding 1e-3.
De-Novo: Trinity/EdgeR
This workflow identified several genes being differentially expressed at a p-value exceeding 1e-3. Annotation of these genes identified them as coming from viral sources.
DE Genes @ P = 1e-3
C820_G1
C10514_G1
C13161_G1
C7917_G1
C11984_G1
C18802_G1
C54_G1
C12935_G1
C476_G1
C14261_G1
C135_G1
933
935
458
460
Example (BLAST):
Conclusion
Quantification methods could not be compared due to the presence of viral sequence data for Drosophila birnavirus and X virus. These genes were not identified using reference-based approach as their source was different from the reference used.
Future Work
➢ Remove virus transcripts at de-novo assembly step and repeat project workflow
➢ Apply draft methods to data from other non-model organism


Methods/Workflow
Data
Retrieve necessary read and annotation data and prepare for assembly.
NGS SRA data
UCSC Reference and Annotation
Improve Data Quality
FastQC, Sickle, Scythe
Sequence Databases
Uniprot
Assembly & Annotation
Use improved data in de-novo and reference based assembly, annotate de-novo assembly for quantification process.
Mapping Assembly
TopHat, Bowtie 1, Cufflinks, Cuffmerge
De-novo Assembly
Trinity
Transcript Annotation
BLAST
Quantification
Produce RPKM / FPKM to be used in differential expression detection process.
Reference Required
Cuffdiff
Reference Free (or Optional)
eXpress, RSEM
Differential Expression
Detect differential expression in samples and visualize via heatmap output.
CummeRbund
EdgeR
DESEQ


Color Key
−2 0 2
Value
samples vs. features
gene_DE_P1e−3_C2.matrix.log2.centered
c820_g1
c10514_g1
c13161_g1
c7917_g1
c11984_g1
c18802_g1
c54_g1
c12935_g1
c476_g1
c14261_g1
c135_g1
rsem_dmel_scytrim933
rsem_dmel_scytrim935
rsem_dmel_scytrim458
rsem_dmel_scytrim460



Example (BLAST): C12935_g1_i1 len = 1061 path = [1039:0-1060]

Top 5 sequences producing significant alignments ordered by coverage:	Max score	Total score	Query cover	E-value	Accession
Drosophila melanogaster birnavirus SW-2009a strain DBV segment A, complete sequence	**448**	**2506**	**99%**	**3e-160**	**GQ342962.1**
Nocardiopsis dassonvillei subsp. Dassonvillei DSM 43111 chromosome 1, complete sequence	42.3	77.3	35%	0.80	CP002040.1
Streptomyces nodosus strain ATCC 14899 genome	39.6	132	35%	7.2	CP009313.1
Actinoplanes sp. N902-109, complete sequence	38.6	74.2	28%	8.4	CP005929.1
PREDICTED: Dasypus novemcinctus ubiquitin specific peptidase 36 (USP36), transcript variant X1, mRNA	S35.0	64.5	27%	7.8	XM_004460442.2

Vo.Bo. _____

Nombre del mentor (utilice letra de molde)

_____

Firma del Mentor

_____

Firma del Estudiante

_____

Fecha

MARC/F-001/0Q